*Published online 14 December 2006* *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D5–D12***
*doi:10.1093/nar/gkl1031*

# Database resources of the National Center for Biotechnology Information

**David L. Wheeler*, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Lewis Y. Geer, Yuri Kapustin, Oleg Khovayko, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, James Ostell, Vadim Miller, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Steven T. Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L. Tatusov, Tatiana A. Tatusova, Lukas Wagner and Eugene Yaschenko**

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 15, 2006; Revised October 16, 2006; Accepted October 17, 2006

## ABSTRACT

**In addition to maintaining the GenBank® nucleic acid sequence database, the National Center for Biotechnology Information (NCBI) provides analysis and retrieval resources for the data in GenBank and other biological data made available through NCBI's Web site. NCBI resources include Entrez, the Entrez Programming Utilities, My NCBI, PubMed, PubMed Central, Entrez Gene, the NCBI Taxonomy Browser, BLAST, BLAST Link(BLink), Electronic PCR, OrfFinder, Spidey, Splign, RefSeq, UniGene, HomoloGene, ProtEST, dbMHC, dbSNP, Cancer Chromosomes, Entrez Genome, Genome Project and related tools, the Trace and Assembly Archives, the Map Viewer, Model Maker, Evidence Viewer, Clusters of Orthologous Groups (COGs), Viral Genotyping Tools, Influenza Viral Resources, HIV-1/Human Protein Interaction Database, Gene Expression Omnibus (GEO), Entrez Probe, GENSAT, Online Mendelian Inheritance in Man (OMIM), Online Mendelian Inheritance in Animals (OMIA), the Molecular Modeling Database (MMDB), the Conserved Domain Database (CDD), the Conserved Domain Architecture Retrieval Tool (CDART) and the PubChem suite of small molecule databases. Augmenting many of the Web applications are custom implementations of the BLAST program optimized to search specialized data sets. These resources can be accessed through the NCBI home page at www.ncbi.nlm.nih.gov.**

## INTRODUCTION

The National Center for Biotechnology Information (NCBI) at the National Institutes of Health was created in 1988 to develop information systems for molecular biology. In addition to maintaining the GenBank® (1) nucleic acid sequence database, to which data is submitted by the scientific community, NCBI provides data retrieval systems and computational resources for the analysis of GenBank data as well as a variety of other biological data. For the purposes of this article, the NCBI suite of database resources is grouped into six broad categories. All resources discussed are available from the NCBI home page at www.ncbi.nlm.nih.gov. In most cases, the data underlying these resources is available for bulk download at ftp.ncbi.nih.gov, a link from the NCBI home page.

## DATABASE RETRIEVAL TOOLS

### Entrez

Entrez (2) is an integrated database retrieval system that enables text searching, using simple Boolean queries, of a diverse set of 31 databases. Global Query, the default search on the NCBI homepage, searches across all the Entrez databases and rapidly returns the counts of matching records in each database. A user may then display results or further refine searches in any individual database. The Entrez databases include ~91 million DNA and protein sequences derived from several sources (1,3–6), the NCBI taxonomy, genomes, population sets, gene expression data, over 1.2 million gene-oriented sequence clusters in UniGene, almost 500 000 sequence-tagged sites in UniSTS, 34 million genetic variations in dbSNP, over 36 000 protein structures

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: wheeler@ncbi.nlm.nih.gov

from the Molecular Modeling Database (MMDB) (6), 168 000 3D and 12 000 alignment-based protein domains, and the biomedical literature via PubMed, Pubmed Central (PMC), Online Mendelian Inheritance in Man (OMIM) and online books. The books database contains >60 online scientific textbooks. To enable researchers to quickly reach the appropriate NCBI resource, the content of the NCBI web pages and FTP directories has been incorporated into an Entrez database of its own. Searches of the NCBI web site using the same powerful queries available for the biological databases are therefore possible.

Entrez provides extensive links within and between database records. In their simplest form, these links may be cross-references between a sequence and the abstract of the paper in which it is reported, or between a protein sequence and its coding DNA sequence or, perhaps, its 3D-structure. Other examples are links between a genomic assembly and its components or between a genomic sequence and those sequences derived from its annotation. Computationally derived links between 'neighboring records' such as those based on computed similarities among sequences or among PubMed abstracts, allow rapid access to groups of related records. A service called LinkOut expands the range of links to include external services, such as organism-specific genome databases. To accommodate the growing number of links, Entrez provides a Links pull down menu that appears in the top, right hand corner of record displays.

The records retrieved in Entrez can be displayed in many formats and downloaded singly or in batches. A redirection control allows results to be saved in a local file, shown in the browser as plain text. Results may also be sent to the Entrez clipboard where they may be recalled later during an Entrez session or saved between sessions using My NCBI, described below. In addition, PubMed results and those from other databases may be emailed directly from Entrez or exported as RSS feeds. Formats available for GenBank records include the GenBank Flatfile, FASTA, XML, ASN.1 and others. Graphical display formats are offered for some types of records, including genomic records. For sequence records, a formatting control allows the display or download of a particular range of residues.

Entrez's 'My NCBI' allows users to store personal configuration options such as search filters, LinkOut preferences and document delivery providers. My NCBI also saves searches and can automatically email updated search results. Entrez uses a set of up to five filter tabs used to display subsets of database results. The tabs vary according to Entrez database; examples of some defaults include 'mRNA' and 'RefSesq' subsets for Nucleotide; a 'Review' subset for PubMed; 'NMR' and 'X-ray' subsets for Structure. Default filter tabs can be changed using My NCBI. Additional My NCBI features include changing the way Entrez links are displayed to standard html links or pull downs, and highlighting PubMed search terms. A recently added My NCBI feature called 'Collections' allows users to save search results and bibliographies indefinitely.

Scripted access to Entrez is provided by the Entrez Programming Utilities (E-Utilities), a suite of eight server-side programs supporting a uniform set of parameters used to search, link between, and download from, the Entrez databases. A search history, available via interactive Entrez as well as via the E-Utilities, allows users to recall the results of previous searches during an Entrez session and combine them using Boolean logic. The 'einfo' utility can be used to retrieve detailed information about the Entrez databases, such as lists of supported search fields or the date of the last database update, while 'egquery' returns the number of matches to a single query in every Entrez database. An automated system may use E-Utilities such as 'efetch' or 'esummary', to retrieve the data. Espell checks spelling within Entrez queries and offers suggestions in cases where a misspelling might cause key records to be missed. Support for the Simple Object Access Protocol (SOAP) interface to the E-Utilities was expanded during the past year, and now supports full downloads (efetch) from nine of the Entrez databases with esearch and esummary, support for all. Instructions for using the E-Utilities are found under the 'Entrez Tools' link on the NCBI home page.

## PubMed and PubMed Central

The PubMed database includes over 16.5 million citations from >19 000 life science journals for biomedical articles back to the 1950s, most with abstracts and many with links to the full-text article. PubMed is heavily linked to other core Entrez databases such as Nucleotide, Protein, Gene, Structure and PubChem where it provides a crucial bridge between the data of molecular biology and the scientific literature. PubMed records are also linked to one another within Entrez as 'related articles' on the basis of computationally detected similarities using indexed Medical Subject Heading (7) terms and the text of titles and abstracts. To put information about the top-ranking related articles at the fingertips of researches, the 'Abstract-Plus' display for single PubMed records was introduced this year as the default format for a single record. Abstract-Plus shows, in addition to the abstract of a paper, succinct descriptions of the top five related articles, increasing the potential for the discovery of important relationships.

PubMed Central (8) is a digital archive of peer-reviewed journals in the life sciences providing access to >750 000 full-text articles, a 50% increase over the past year. More than 270 journals, including Nucleic Acids Research, deposit the full text of their articles in PMC. It includes digitized back content for many journals, going back in some cases to the 1800s or early 1900s. Participation in PMC requires a commitment to free access to full text, either immediately after publication or within a 12month period. All PMC free articles are identified in PubMed search results and PMC itself can be searched using Entrez.

## Taxonomy

The NCBI taxonomy database, growing at the rate of 2900 new taxa a month, indexes >240 000 named organisms that are represented in the databases with at least one nucleotide or protein sequence. The Taxonomy Browser can be used to view the taxonomic position or retrieve data from any of the principal Entrez databases for a particular organism or group. The Taxonomy Browser also displays links to the Map Viewer, Genomic BLAST services, the Trace Archive, and to external model organism and taxonomic databases via LinkOut. Searches of the NCBI taxonomy may be made

on the basis of whole, partial or phonetically spelled organism names. Entrez Taxonomy displays include custom taxonomic trees representing user-specified subsets of the full NCBI taxonomy.

## THE BLAST FAMILY OF SEQUENCE-SIMILARITY SEARCH PROGRAMS

The BLAST programs (9–11) perform sequence-similarity searches against a variety of databases, returning a set of gapped alignments with links to full database records, to UniGene, Gene, the MMDB or Gene Expression Omnibus (GEO). One variant, BLAST2Sequences (12), compares two DNA or protein sequences and produces a dot-plot representation of the alignments. In addition to the Web based versions described below, the basic BLAST programs are available as standalone commandline programs, as network clients and as a local web-server package (ftp://ftp.ncbi.nih.gov/blast/executables/LATEST/).

Each alignment returned by BLAST is scored and assigned a measure of statistical significance called the Expectation Value (*E*-value). BLAST takes into account the amino-acid composition of the query sequence in its estimation of statistical significance. This composition-based statistical treatment, used in conventional protein BLAST searches as well as PSI-BLAST searches, tends to reduce the number of false-positive database hits (13). The alignments returned can be limited by an *E*-value threshold or range.

### BLAST output formats

Standard output formats include the default pairwise alignment, several query-anchored multiple sequence alignment formats, an easily parsable Hit Table and a taxonomically organized output. Database sequences appearing in BLAST results may be marked for batch retrieval using check boxes. Alignments against database sequences that are >200 000 bp in length are displayed with links to nearby features, such as genes. A 'Pairwise with identities' mode better highlights differences between the query and a target sequence. An option to display masked characters in lowercase or using distinct colors is also available. A new Tree View option for the Web BLAST service creates a dendrogram that clusters sequences according to their distances from the query sequence. This display is helpful for recognizing the presence of aberrant or unusual sequences or natural groupings of related sequences such as members of a gene family or homologs from other species in the BLAST output.

### Web BLAST

A powerful feature of the NCBI Web BLAST interface is that it allows both the initial search and the results displayed to be restricted to a database subset using an Entrez query as a filter. Web BLAST also uses a standard URL-API that allows complete search specifications, including BLAST parameters, such as Entrez restrictions and the search query, to be contained in a URL posted to the web page.

### MegaBLAST

MegaBLAST (14), designed to find nearly exact matches, is available through a Web interface that handles batch nucleotide queries and operates up to 10 times faster than standard nucleotide BLAST. MegaBLAST is the default search program for NCBI's Genomic BLAST pages. MegaBLAST is also used to search the rapidly growing Trace Archive and is available for the standard BLAST databases as well. For rapid cross-species nucleotide queries, NCBI offers Discontiguous MegaBLAST which uses a non-contiguous word match (15) as the nucleus for its alignments. Discontiguous MegaBLAST is far more rapid than a translated search such as blastx, yet maintains a competitive degree of sensitivity when comparing coding regions.

### Genomic BLAST

NCBI maintains Genomic BLAST pages for >45 organisms shown in the Map Viewer. Genomic BLAST may be used to search the genomic sequence of an organism, the nucleotide and protein Reference Sequences (RefSeqs) used in, and resulting from, the annotation of the genomic sequence, or sets of sequences, such as expressed sequence tags (ESTs), that are mapped to the genomic sequence. The results of Genomic BLAST searches can be displayed within their genomic context using the Map Viewer to show, in addition to the location of a match to the query sequence, the locations of neighboring genes and nearby genomic landmarks. During the past year, a system was implemented to generate these BLAST pages automatically, allowing new organisms to be added very quickly and providing a higher degree of uniformity in database content.

## RESOURCES FOR GENE-LEVEL SEQUENCES

### Databases

*Entrez Gene*. Entrez Gene (16), provides an interface to curated sequences and descriptive information about genes with links to NCBI's Map Viewer, Evidence Viewer (EV), Model Maker (MM), BLAST Link (BLink), protein domains from the Conserved Domain Database (CDD) and other gene-related resources. Data is accumulated and maintained through several international collaborations in addition to curation by in-house staff. Links within Gene to the newest citations in PubMed are maintained by curators and provided as Gene References into Function (GeneRIF). The GeneRIF link within Gene reports leads to a form allowing researchers to add GeneRIFs to a Gene report. Entrez Gene displays use a collapsible navigation panel containing a table of contents for the record, the set of links to other resources, and links to related NCBI tools. The complete Entrez Gene data set, as well as organism-specific subsets, is available in the compact NCBI ASN.1 format on the NCBI FTP site. A tool that converts the native Gene ASN.1 format into XML, called 'gene2xml' is available for several popular computer platforms at ftp.ncbi.nih.gov/toolbox/ncbi\s\do5(t)ools/converters/by\s\do5(p)rogram/gene2xml/. The tool supports filtering by organism so that organism-specific XML files can easily be generated from the comprehensive ASN.1 FTP file.

*UniGene*. UniGene (17), is a system for partitioning GenBank sequences, including ESTs, into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, and is linked to the tissue

types in which the gene is expressed, model organism protein similarities and Entrez Gene. UniGene clusters are created for all organisms for which there are 70 000 or more ESTs in GenBank and includes ESTs for some 37 animals and another 37 plants and fungi. For the human UniGene June 2006 release (build 194), over 7.0 million human ESTs in GenBank were reduced 80-fold in number to ~87 000 sequence clusters. When sufficient genomic sequence is available, UniGene clusters are built using a genome-based clustering system to identify sets of transcript sequences which correspond to distinct transcription loci or to annotated genes. The procedure used for genome-based clustering of transcript sequences is described at http://www.ncbi.nlm.nih.gov/UniGene/build2.html. The UniGene collection has been used as a source of unique sequences for the fabrication of microarrays for the large-scale study of gene expression (18). UniGene databases are updated weekly with new EST sequences, and bimonthly with newly characterized sequences.

*ProtEST*. ProtEST, tightly coupled to UniGene, presents pre-computed BLAST alignments between protein sequences from model organisms and the six-frame translations of nucleotide sequences in UniGene. ProtEST links are displayed in UniGene reports with model organism protein similarities.

*HomoloGene*. HomoloGene is a system for automated detection of homologs among the annotated genes of 18 completely sequenced eukaryotic genomes including those of *Homo sapiens, Pan troglodytes, Mus musculus, Rattus norvegicus, Drosophila melanogaster, Anopheles gambiae, Caenorhabditis elegans, Schizosaccharomyces pombe, Saccharomyces cerevisiae, Eremothecium gossypii, Neurospora crassa, Magnaporthe grisea, Arabidopsis thaliana* and *Oryza sativa*. The HomoloGene build procedure is guided by the taxonomic tree, and relies on conserved gene order and measures of DNA similarity among closely related species, while making use of protein similarity for more distantly related organisms. HomoloGene reports include homology and phenotype information drawn from OMIM, Mouse Genome Informatics (19), Zebrafish Information Network (20), Saccharomyces Genome Database (21), Clusters of Orthologous Groups (COG) (22) and FlyBase. A Pairwise Scores display gives a table of statistics for protein and nucleotide sequences among members of a Homologene group. HomoloGene entries include paralogs in addition to orthologs. HomoloGene can be queried using Entrez at www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene.

*The Database for the Major Histocompatbility Complex*. The Database for the Major Histocompatbility Complex (dbMHC) contains variations found in alleles of the Major Histocompatibility Complex (MHC), a highly variable array of genes playing a vital role in the success of organ transplants and susceptibility to infectious diseases. dbMHC contains hundreds of sequences for MHC alleles and data on typing kits used by academic, clinical and industrial laboratories. The database includes data arising from a survey of Human Leukocyte Antigen (HLA) allele frequency distributions as well as a project to collect HLA genotype and clinical outcome information on hematopoietic cell transplants performed worldwide. In addition, support for three new projects related to Type 1 Diabetes, Rheumatoid Arthritis and Natural Killer Cell Immunoglobulin-like Receptors, respectively, was added over the past year. Access to the data, lists of contributors, as well as a number of online tools for data analysis are provided at www.ncbi.nlm.nih.gov/mhc/MHC.cgi?cmd=init.

*A database of single nucleotide polymorphisms (dbSNP)* . The database of single nucleotide polymorphisms (23), a repository for single base nucleotide substitutions and short deletion and insertion polymorphisms, contains over 12 million human SNPs and another 22 million from a variety of other organisms, with 14 million of these added over the past year. SNP reports link to 3D structures from the MMDB via NCBI's interactive macromolecular structure viewer, Cn3D (24), to highlight implied amino acid changes in coding regions. dbSNP provides additional information about the validation status, population-specific allele frequencies and individual genotypes for dbSNP submission. These data are available on the dbSNP FTP site in XML-structured genotype reports that include information and about cell lines, pedigree IDs and error flags for genotype inconsistencies and incompatibilities. Haplotype and linkage disequilibrium data are being incorporated in dbSNP as data are released from the International HapMap project. Functional variants are identified when dbSNP submissions can be matched to OMIM records and mutation reports in the biomedical literature. Entrez SNP supports searches for SNPs lying between two markers and batch downloads.

*Reference sequences (RefSeq)*. The Reference Sequence (RefSeq) database (3), provides curated references for transcripts, proteins and genomic regions, plus computationally derived nucleotide sequences and proteins. The complete RefSeq database is provided in the RefSeq directory on the NCBI FTP site. The number of sequences in RefSeq has grown by 46% over the past year. As of Release 18, RefSeq contained over 4.1 million sequences, including >2.8 million protein sequences, representing almost 3695 organisms.

## Tools for gene-level analysis

*Open reading frame finder (ORF)*. ORF finder performs a six-frame translation of a nucleotide sequence and returns the location of each ORF within a specified size range. Translations of the ORFs detected can be analyzed via BLAST against the standard BLAST or COGs databases.

*Splign and Spidey*. Splign (25) is a utility for computing cDNA-to-genomic, or spliced sequence alignments that is accurate in determining splice sites, tolerant of sequencing errors and supports cross-species alignments. Splign uses a version of the Needleman–Wunsch algorithm (26) that accounts for splice signals in combination with BLAST to identify possible locations of genes and their copies as well as to speed up the core dynamic programming. The web version of Splign is able to compute and display the spliced alignment of a transcript sequence to a genomic sequence of up to 50 Mb in seconds. A standalone version that operates on longer genomic sequences is found at www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi?textpage=downloads.

Spidey aligns a set of eukaryotic mRNA sequences to against a single genomic sequence taking into account predicted splice sites and using one of four splice-site models

(Vertebrate, *Drosophila*, *C.elegans*, Plant). Spidey returns exon alignments, protein translations and a summary showing the alignment quality and goodness of match to splice junction patterns for each putative exon.

*Electronic PCR*. Two types of Electronic PCR (e-PCR) can be performed from the e-PCR home page at www.ncbi.nlm.nih.gov/sutils/e-pcr. Forward e-PCR searches for matches to STS primer pairs in the UniSTS database of almost unistsnumber markers. Reverse e-PCR is used to estimate the genomic binding site, amplicon size and specificity for sets of primer pairs by searching against the genomic and transcript databases of *A.gambiae, A.thaliana, C.elegans, Danio rerio, D.melanogaster, H.sapiens, M.musculus* and *R.norvegicus*. To increase sensitivity, Forward e-PCR allows the size of the primer segment to be matched, and the number of mismatches, number of gaps and the size of the STS to be adjusted. Binaries for several computer platforms, along with the source code, are available via FTP at ftp.ncbi.nlm.nih.gov/pub/schuler/e-PCR.

Web interfaces to OrfFinder, Spidey, Splign and e-PCR are available via the 'Tools' link on the NCBI home page.

## RESOURCES FOR GENOME-SCALE ANALYSIS

### Databases for genomic analysis

*Entrez Genome*. Entrez Genome (27) provides access to over 370 complete microbial genomic sequences (120 added over the past year), >2450 viral genomic sequences (350 added) and >1050 reference sequences for eukaryotic organelles (250 added). Over 20 higher eukaryotic genomes are also included, such as the recent arrival, *Macaca mulatta,* the rhesus monkey. The Plant Genomes Central web page serves as a portal to completed plant genomes, to information on plant genome sequencing projects, or to plant-related resources at NCBI such as plant Genomic BLAST pages or Map Viewer. Specialized viewers and BLAST pages, are also available for eukaryotic organelles and viruses.

Within Entrez Genomes, genomes are chosen from an alphabetical listing or a phylogenetic tree and can be examined at increasing levels of detail ranging from a graphical overview of an entire genome to the level of a single gene. At the level of a genome or a chromosome, a Coding Regions display gives the locations coding regions, and the lengths, names and GenBank identification numbers of the protein products. An RNA Genes view lists the location and names for ribosomal and transfer RNA genes. A summary of COG functional groups is also presented. At the level of a single gene, links are provided to sequence neighbors for the implied protein with links to the COGs database.

For complete microbial genomes, pre-computed BLAST neighbors for protein sequences, including their taxonomic distribution and links to 3D structures, are given in TaxTables and PDBTables, respectively. Pairwise sequence alignments are linked to the Cn3D macromolecular structure viewer (24) to generate displays of 3D structures coupled to sequence alignments. The TaxPlot, GenePlot and gMap tools, described below, are available as links from Genomes displays.

*The trace and assembly archives*. The Trace Archive is a rapidly growing database of >1.3 billion sequencing traces. More than 860 organisms are represented, an increase of >100 over the past year. The Assembly Archive links the raw sequence information found in the Trace Archive with assembly information found in GenBank. An Assembly Viewer allows displays of multiple sequence alignments as well as the sequence chromatograms for traces that are part of assemblies. The Trace Assembly Archives are linked from the NCBI home page.

*Genome Project*. The Entrez Genome Project database supplements Entrez Genome by providing an overview of the status of complete and in-progress large-scale sequencing, assembly, annotation and mapping projects. Genome Project links to project data in the other Entrez databases, such as Entrez Nucleotide and Genome, and to a variety of other NCBI and external resources. For prokaryotic organisms, Genome Project indexes a number of characteristics of interest to biologists such as organism morphology and motility; environmental requirements, such as salinity, temperature and pH range; oxygen requirements and pathogenicity. The database allows genome sequencing centers to register their project early in the sequencing process so that project data can be linked to other NCBI-hosted data at the earliest opportunity.

### Other resources for genomic analysis

*Map Viewer*. The NCBI Map Viewer displays genome assemblies, genetic and physical markers and the results of annotation and other analyses using sets of aligned maps. The Map Viewer home page (www.ncbi.nlm.nih.gov/mapview/) provides links to both Map Viewer and Genomic BLAST pages from a taxonomically organized organism list of over 45 organisms including *H.sapiens, M.musculus* and *R.norvegicus*. Maps available for display in the Map Viewer vary by organism but may include cytogenetic maps, physical maps, maps showing predicted gene models, EST alignments with links to UniGene clusters, and mRNA alignments used to construct gene models. Maps from multiple organisms or multiple assemblies for the same organism can be displayed in a single view. A feature recently added to the Map Viewer is the ability to view previous genome builds. The Map Viewer supports queries using various identifiers such as gene names or symbols, marker names, SNP identifiers, or accession numbers. Plant genomes in the Map Viewer can be queried in tandem using a cross species query page to generate a display of the chromosome maps from multiple species. The Map Viewer can generate a tabular display for convenient export to other programs and segments of a genomic assembly may be downloaded using a Download/View Sequence link. Map Viewer displays link to Entrez Gene, and to tools such as the Evidence Viewer and Model Maker. Map Viewer links in the Entrez Links menu for Gene, Nucleotide or Protein databases provide a convenient route to display a region of interest.

*Model Maker*. Model Maker (MM) is used to construct transcript models using combinations of putative exons derived from *ab initio* predictions or from the alignment of GenBank transcripts, including ESTs, and RefSeqs, to the NCBI human genome assembly. Previously observed exon splice patterns

are indicated as guides to model building. Completed models may be saved locally or analyzed with OrfFinder.

*Evidence Viewer*. The Evidence Viewer (EV) displays the alignments to genomic contigs of RefSeq and GenBank transcripts, and ESTs supporting gene models. Mismatches between transcript and genomic sequences are highlighted. Exon-by-exon transcript alignments, including flanking genomic sequence for each exon, are given along with protein translations. Proteins annotated on the transcript sequences are shown and mismatches between proteins annotated on the aligned transcripts are highlighted.

*Cancer Chromosomes*. Three databases, the NCI/NCBI SKY (Spectral Karyotyping)/M-FISH (Multiplex-FISH) and CGH (Comparative Genomic Hybridization) Database, the NCI Mitelman Database of Chromosome Aberrations in Cancer (28), and the NCI Recurrent Chromosome Aberrations in Cancer databases comprise the new Cancerchromosomes Entrez database found at www.ncbi.nlm.nih.gov/entrez/query.fcgi? db=CancerChromosomes.

Three search formats are available: a conventional Entrez query, a Quick/Simple search, and an Advanced Search. The simple search offers a set of menus used to select a disease site or diagnosis that can be combined with specifications for a particular chromosomal location and anomaly. The Advanced Search offers a combination of forms for more complex queries. Search results may list all cases matching the query terms, a case-based report, or list each clone or cell separately, the clone/cell report. Similarity reports show terms common to a group or records within several term categories, such as diagnosis or disease site and cytogenetic abnormalities, among the selected cases or clones/cells.

*TaxPlot, GenePlot and gMap*. TaxPlot plots similarities in the proteomes of two organisms to that of a reference organism for >580 prokaryotic and ~50 eukaryotic genomes. A related tool, GenePlot, generates plots of protein similarity for a pair of complete microbial genomes to visualize deleted, transposed or inverted genomic segments.

The 'gMap' tool combines the results of pre-computed whole microbial genome comparisons with on-the-fly BLAST comparisons clustering genomes with similar nucleotide sequences and then represents the pre-computed segments of similarity graphically. A novel sequence may be introduced into the display of pre-computed alignments using an on-the-fly BLAST comparison. Using gMap, one can quickly navigate from a high-level overview of the similarity between a set of genomes, to a nucleotide-level view of individual segments of alignment.

*Clusters of orthologous groups (COGs)*. The rapid progress in sequencing has produced sequences for >370 prokaryotic genomes comprising >250 species within 145 different taxonomic genera. The COGs database (22), presents a compilation of orthologous groups of proteins from completely sequenced organisms. A eukaryotic version, KOGs, is available for seven organisms including *H.sapiens; C.elegans, D.melanogaster* and *A.thaliana*. Alignments of sequence from COGs have been incorporated into the CDD described below.

*Viral genotyping tools*. NCBI offers a web-based genotyping tool that employs a blastn comparison between a viral sequence to be subtyped and either a default panel of reference sequences or a panel provided by the user.

*Influenza genome resources*. The Influenza Genome Sequencing Project (IGSP) (29) is providing researchers with a growing collection of virus sequences essential to the identification of the genetic determinants of influenza pathogenicity. To date, the project has generated ~12 000 influenza sequences. NCBI's new Influenza Virus Resource links the IGSP project data, via PubMed, to the most recent scientific literature on influenza as well as to a number of on line analysis tools and databases. These databases include NCBI's Influenza Virus Sequence Database, comprised of ~34 000 influenza sequences in GenBank, and NCBI's RefSeq database. Using the tools of the Influenza Virus Resource, researchers can extend their analyses to the 42 000 influenza protein sequences, 101 influenza protein structures and 240 influenza population studies accessible within the biological databases covered by NCBI's Entrez system. A recently introduced online influenza genome annotation tool analyzes a novel sequence and produces output in a 'feature table' format that can be used by NCBI's GenBank submission tools such as 'tbl2asn' (1).

## RESOURCES FOR THE ANALYSIS OF PATTERNS OF GENE EXPRESSION AND PHENOTYPES

### Resources for the analysis of gene expression

*Gene expression omnibus (GEO)*. The Gene Expression Omnibus (GEO) (30) is a data repository and retrieval system for microarray and other forms of high-throughput molecular abundance data generated by the scientific community. In addition to gene expression data, GEO accepts array comparative genomic hybridization (aCGH) data, chromatin immunoprecipitation on array (ChIP-chip) data, SNP array data and some proteomic data types. The GEO repository accepts Minimum Information About a Microarray Experiment (MIAME)-compliant data submissions. Several data deposit options and formats are supported, including web forms, spreadsheets, XML and Simple Omnibus Format in Text (SOFT). The repository may be queried from both experiment—(Entrez GEO DataSets) and gene-centric (Entrez GEO Profiles) perspectives. At the time of writing, the repository contains data from ~120 000 hybridization experiments, representing ~3 billion individual measurements, derived from ~3000 array definitions, and spanning >200 organisms.

*GENSAT*. GENSAT is a gene expression atlas of the mouse central nervous system produced with data supplied by the National Institute of Neurological Disorders and Stroke. GENSAT catalogs images of histological sections of the mouse brain in which tags, such as enhanced green fluorescence protein, have been used to visualize the relative degree of localized expression for a wide array of genes. Images are available for the mouse brain at various developmental stages. GENSAT records link to to Entrez Gene, UniGene, GEO, PubMed and PubMed Central.

*Probe*. Nucleic acid probes are molecules that complement a specific gene transcript or DNA sequence and are useful in gene silencing, genome mapping and genome variation analysis. The new Entrez Probedatabase serves as an archive

of probe sequences along with data on their experimental utility. Probe entries indicate the intended experimental application and include the experimental results generated using the probe. Entrez Probe is linked to the scientific literature in PubMed as well as to Entrez Nucleotide with pre-computed alignments to RefSeqs providing a bridge to the genomic information in Entrez Gene.

### Resources for the analysis of phenotypes

*Online Mendelian Inheritance in Man (OMIM).* NCBI provides the online version of the OMIM catalog of human genes and genetic disorders authored and edited by Victor A. McKusick at The Johns Hopkins University (31,32). The database contains information on disease phenotypes and genes, including extensive descriptions, gene names, inheritance patterns, map locations, gene polymorphisms and detailed bibliographies. The OMIM Entrez database contains ~17 000 entries, including data on >11 000 established gene loci and phenotypic descriptions. These records link many important resources, such as locus-specific databases and GeneTests.

*Online Mendelian Inheritance in Animals (OMIA).* Online Mendelian Inheritance in Animals (OMIA) is a database of genes, inherited disorders and traits in animal species, other than human and mouse, authored by Professor Frank Nicholas of the University of Sydney, Australia, and colleagues. The database contains textual information and references, as well as links to relevant records from OMIM, PubMed, Entrez Gene.

## RESOURCES FOR MOLECULAR STRUCTURE AND PROTEOMICS

### Structure databases

*The molecular modeling database.* The NCBI MMDB, built by processing entries from the Protein Data Bank (5), is described in (33). The structures in the MMDB are linked to sequences in Entrez and to the Conserved Domain Database (6).

*The conserved domain database and conserved domain architecture retrieval tool.* The CDD contains >12 000 PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the Simple Modular Architecture Research Tool (SMART) (34), Pfam (35), and from domain alignments derived from COGs. NCBI's Conserved Domain Search (CD-Search) service can be used to search a protein sequence for conserved domains in the CDD. Wherever possible CDD hits are linked to structures which, coupled with a multiple sequence alignment of representatives of the domain hit, can be viewed with NCBI's 3D molecular structure viewer, Cn3D (24), equipped with advanced alignment-building tools that use the PSI-BLAST and threading algorithms. The Conserved Domain Architecture Retrieval Tool (CDART) allows searches of protein databases on the basis of a conserved domain and returns the domain architectures of database proteins containing the query domain. Alignment-based protein domain information from the CDD and 3D domains from the MMDB can be searched via the Entrez interface.

*PubChem.* PubChem is the informatics backbone for the NIH Roadmap Initiative on molecular libraries. PubChem focuses on the chemical, structural and biological properties of small molecules, particularly their application as diagnostic and therapeutic agents. A suite of three Entrez databases, PCSubstance, PCCompound and PCBioAssay, debuted during the past year to contain the substance information, compound structures and bioactivity data of the PubChem project. The databases comprise records for over 12.8 million compounds with over 8 million unique structures. The PubChem databases link to other Entrez databases such as PubMed and PubMed Central but also to Entrez Structure and Protein to provide a bridge between the macromolecules of genomics and the small organic molecules of cellular metabolism.

### Tools supporting proteomics

*Blast Link (BLink).* BLink displays pre-computed BLAST alignments to similar sequences for each protein sequence in the Entrez databases. BLink can display alignment subsets limited by taxonomic criteria, by database of origin, relation to a complete genome, membership in a COG (22) or by relation to a 3D structure or conserved protein domain. BLink links are displayed for protein records in Entrez as well as within Entrez Gene reports.

*The open mass spectrometry search algorithm.* The Open Mass Spectrometry Search Algorithm (OMSSA) (36) is an efficient search engine for identifying MS/MS peptide spectra by searching libraries of known protein sequences. OMSSA assigns significant hits a Expect-value computed in the same way as the *E*-value of BLAST. The web interface to OMSSA, reached via a link from the 'Tools' link on the NCBI home page, allows up to 2000 spectra to be analyzed in a single session using either the BLAST 'nr' or 'refseq_protein' sequence libraries for comparison. Standalone versions of OMSSA for several popular computer platforms that accept larger batches of spectra and allow searches of custom sequence libraries can be downloaded at pubchem.ncbi.nlm.nih.gov/omssa/download.htm.

*HIV-1/Human protein interaction database.* The Division of Acquired Immunodeficiency Syndrome of the National Institute of Allergy and Infectious Diseases, in collaboration with the Southern Research Institute and NCBI, maintains a comprehensive HIV Protein-Interaction Database of documented interactions between HIV-1 proteins, host cell proteins, other HIV-1 proteins, or proteins from disease organisms associated with HIV or AIDS. Summaries, including protein RefSeq accession numbers, Entrez Gene IDs, lists of interacting amino acids, brief descriptions of interactions, keywords and PubMed IDs for supporting journal articles are presented at www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/index.html. Interaction summaries are selected using pull down phrase lists to apply filters, and batches of summaries may be downloaded. All protein–protein interactions documented in the HIV Protein-Interaction Database are listed in Entrez Gene reports in the HIV-1 protein interactions section.

## SOURCES OF FURTHER INFORMATION

The resources described here include documentation, other explanatory material and references to collaborators and data sources on the respective web sites. The NCBI

Handbook, available in the Books database, describes the principal NCBI resources in detail. Several tutorials are also offered under the Education link from NCBI's home page. A Site Map provides a comprehensive table of NCBI resources, and the About NCBI feature provides bioinformatics primers and other supplementary information. A user support staff is available to answer questions at info@ncbi.nlm.nih.gov. Updates on NCBI resources and database enhancements are described in the NCBI News newsletter (http://www.ncbi.nlm.nih.gov/About/newsletter.html). In addition, a number of mailing lists provide updates on resources such as BLAST, Books, the Entrez Utilities, Gene, Genomes, LinkOut, RefSeq, dbSNP and others (http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). RSS feeds for some NCBI resources (http://www.ncbi.nlm.nih.gov/feed/) are also now available.

## ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) GenBank. *Nucleic Acids Res*., **34**, D16–D20.
2. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol*., **266**, 141–162.
3. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.
4. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
5. Kouranov,A., Xie,L., de la Cruz,J., Chen,L., Westbrook,J., Bourne,P.E. and Berman,H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res*., **34**, D302–D305.
6. Marchler-Bauer,A., Anderson,J.B., Cherukuri,P.F., DeWeese-Scott,C., Geer,L.Y., Gwadz,M., He,S., Hurwitz,D.I., Jackson,J.D., Ke,Z. *et al*. (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res*., **33**, D192–D196.
7. Sewell,W. (1964) Medical Subject Headings in MEDLARS. *Bull Med. Libr. Assoc*., **52**, 164–170.
8. Sequeira,E. (2003) PubMed central—three years old and growing stronger. *ARL*, **228**, 5–9.
9. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
10. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
11. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res*., **34**, W6–W9.
12. Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett*., **174**, 247–250.
13. Schäffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*., **29**, 2994–3005.
14. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol*., **7**, 203–214.
15. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
16. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.,* this Issue.
17. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med*., **75**, 694–698.
18. Ermolaeva,O., Rastogi,M., Pruitt,K.D., Schuler,G.D., Bittner,M.L., Chen,Y., Simon,R., Meltzer,P., Trent,J.M. and Boguski,M.S. (1998) Data management and analysis for gene expression arrays. *Nature Genet*., **20**, 19–23.
19. Blake,J.A., Eppig,J.T., Bult,C.J., Kadin,J.A. and Richardson,J.E. (2006) The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res*., **34**, D562–D567.
20. Sprague,J., Bayraktaroglu,L., Clements,D., Conlin,T., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Mani,P., Ramachandran,S. *et al*. (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res*., **34**, D581–D585.
21. Hirschman,J.E., Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G., Hong,E.L., Livstone,M.S., Nash,R. *et al*. (2006) Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res*., **34**, D442–D445.
22. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
23. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.
24. Wang,Y., Geer,L.Y., Chappey,C., Kans,J.A. and Bryant,S.H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci*., **25**, 300–302.
25. Kapustin,Y., Souvorov,A. and Tatusova,T. (2004) Splign—a hybrid approach to spliced alignments. In Gramada,A. and Bourne,P. (eds), *Eighth Annual International Conference on RECOMB 2004—Currents in Computational Molecular Biol.,* p. 741.
26. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*., **48**, 443–453.
27. Tatusova,T.A., Karsch-Mizrachi,I. and Ostell,J.A. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
28. Mitelman,F., Mertens,F. and Johansson,B. (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet*., **15**, 417–474.
29. Ghedin,E., Sengamalay,N.A., Shumway,M., Zaborsky,J., Feldblyum,T., Subbu,V., Spiro,D.J., Sitz,J., Koo,H., Bolotov,P. *et al*. (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, **437**, 1162–1166.
30. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: Mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.,* this Issue (Database Issue).
31. McKusick,V.A. (1998) *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders*. The Johns Hopkins University Press, Baltimore.
32. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **33**, D514–D517.
33. Chen,J., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al*. (2003) MMDB: Entrez's 3D-structure database. *Nucleic Acids Res*., **31**, 474–477.
34. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*., **34**, D257–D260.
35. Finn,R.D., Mistry,J., Schuster-Bo?ckler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res*., **34**, D247–D251.
36. Geer,L.Y., Markey,S.P., Kowalak,J.A., Wagner,L., Xu,M., Maynard,D.M., Yang,X., Shi,W. and Bryant,S.H. (2004) Open mass spectrometry search algorithm. *J. Proteome Res*., **3**, 958–964.